



November 22, 2024

**VIA ECF**

Honorable Ona T. Wang  
United States District Court  
500 Pearl Street  
New York, NY 10007

Re: *Authors Guild v. OpenAI Inc.*, No. 23-cv-8292 and *Alter v. Open AI Inc.*, 23-cv-10211  
(S.D.N.Y.): Response to Brief Regarding Microsoft's LLMs and Licensing (Dkt. 269)

Dear Magistrate Judge Wang:

There can be no dispute that Class Plaintiffs' claims only relate to the training of *OpenAI's* LLMs. There is also no dispute that OpenAI trained those models and then licensed them to Microsoft (a separate company) for inclusion into Microsoft's products. As is publicly known, Microsoft is also building its own LLMs but those are not alleged in the Complaint. With this background, Microsoft opposes Class Plaintiffs' request that the Court compel it to disclose (1) "documents relating to Microsoft's efforts to license LLM training data where a license did not result" as well as (2) "documents showing the datasets that Microsoft used to train its LLMs." (Dkt. 269 at 1.) For the reasons below, the Court should deny this expansion of discovery into models, products, and issues that are not connected to Class Plaintiffs' allegations.

**Because Microsoft's LLMs Are Irrelevant, So are Documents Related to Unexecuted Data Access Agreements.** Class Plaintiffs seek wide-ranging discovery into Microsoft's data access strategy, which has nothing to do with OpenAI's LLMs at issue in this case. This is improper. There is no allegation that Microsoft is sharing with OpenAI the data that it is discussing with third parties. As such, Microsoft's discussions and negotiations regarding data access agreements are neither relevant nor proportional to the needs of this case. Wishing to avoid unnecessary discovery battles, Microsoft agreed to produce the data access agreements that it has actually concluded with third parties, as such evidence could possibly be relevant to the question of whether there is a market for data irrespective of the model at issue. Certainly, this does not mean that intrusive discovery into Microsoft's highly sensitive negotiations and discussions regarding potential future data access agreements and agreements that never came to fruition are relevant in a case that has nothing to do with Microsoft's own model development.

In any event, Microsoft's voluntary production related to executed data access agreements already provides Class Plaintiffs with any relevant discovery. The executed data access agreements may be relevant to whether there is an actual or potential market for training data as they represent a meeting of the minds, which is necessarily not present for unexecuted agreements. To the extent that licensing is even relevant to willfulness,<sup>1</sup> which Class Plaintiffs use again as a catchall justification, mere negotiations or discussions between Microsoft and unrelated third parties

---

<sup>1</sup> *Broadcast Music, Inc. v. Prana Hospitality, Inc.*, discusses willfulness in the context of whether the defendant knew that it was required to take a license for the allegedly infringing conduct. 158 F. Supp. 3d 184, 197-198 (S.D.N.Y. 2016). Here the data access agreements do not relate to the training of the GPT models at issue in this case.

regarding access to data say nothing about whether Microsoft thought a license for training was necessary as to the Class Plaintiffs. As to damages, none of the cases Class Plaintiffs cite state that negotiations regarding *unexecuted* data access agreements involving the *defendant* and unrelated to the copyrighted works-in-suit are relevant. While *On Davis v. The Gap* discusses the “fair market value of a reasonable license,” 246 F.3d 152, 166 (2d Cir. 2001), the only licenses reflecting the true fair market value are executed licenses. Again, negotiations or discussion about unexecuted agreements cannot show a fair market value because there was no actual agreement on the value. The remaining damages cases discuss the relevance of *plaintiff’s* licensing negotiations and, specifically, demands or offers for the copyrighted works-in-suit. See *Ringgold v. Black Entm’t TV, Inc.*, 126 F.3d 70, 81 (2d Cir. 1997)<sup>2</sup>; *Reilly v. Commerce*, 2016 WL 6837895, at \*9 (S.D.N.Y. Oct. 31, 2016). As such, Class Plaintiffs failed to meet their burden of showing the relevance of negotiations about unexecuted agreements for works other than the works-in-suit.

Class Plaintiffs’ reliance on OpenAI’s decision to produce negotiations and discussions regarding its own unexecuted data access agreements is no proxy for this discovery directed to Microsoft. Microsoft and OpenAI are different companies with different product offerings. As such, with respect to Microsoft (whose own LLMs are not at issue in this case), it makes sense to draw the line to only executed data access agreements, as Microsoft has already agreed to produce.

**The Datasets Used to Train Microsoft’s LLMs are Equally Irrelevant.** Discovery into the datasets Microsoft used to train its own LLMs, which are not at issue in this case, is improper. The Complaint only alleges infringement by training of *OpenAI’s LLMs*; and the content of datasets used to train *Microsoft’s LLMs* has no bearing on Microsoft’s knowledge or willfulness with respect to the training of OpenAI’s LLMs. Nor could it. This is nothing more than a fishing expedition aimed at expanding this case well beyond the allegations in the Complaint.

Microsoft’s models are not now and never have been part of Class Plaintiffs’ claims. Class Plaintiffs cherry-pick broad conclusory allegations in their Complaint to argue that models trained by Microsoft—but not named in the Complaint—are relevant to their claims. But the Complaint does not include *any* allegations about those alleged models, let alone allegations that meet the pleading requirements for claims arising from them. The Complaint in this case shows that Plaintiffs’ allegations focus solely on GPT models that were trained by OpenAI, and products that use those models, including ChatGPT and Microsoft’s Copilot.<sup>3</sup> The *only* alleged harm to authors derives from “ChatGPT and the LLMs underlying it.” (Compl. ¶¶ 142-163.) The *only* alleged source of commercial benefit to Microsoft is “GPT-based commercial offerings.” (*Id.* ¶¶ 164-167.) The *only* alleged exploitation of Plaintiffs’ copyrighted works is through the training of GPT models.<sup>4</sup> Even the allegations related to Microsoft for jurisdiction and venue are based on GPT-based products (*Id.* ¶¶ 15, 17.) In fact, when Class Plaintiffs describe Microsoft’s relevance to this lawsuit, they do so in terms of how Microsoft allegedly assisted OpenAI’s training of OpenAI’s GPT models: “In the course of designing and maintaining these tailored supercomputing systems *for OpenAI’s needs*, upon information and belief, *Microsoft was both directly involved in making reproduction of copyrighted material and facilitated the copyright infringement committed by*

<sup>2</sup> The cited discussion in *Ringgold* does not address damages, but rather the fourth fair use factor. 126 F.3d at 80-81.

<sup>3</sup> See, e.g., Compl. ¶¶ 4, 6, 9, 14-15, 17, 52, 83-167, 169, 176-180, 186-190, 196-199, 205-209, 215-219, 225-229, 235-239, 244-247, 253-257, 263-265, 270-274, 279-285, 291-295, 301-305, 311-315, 321-325, 331-335, 341-347, 349, 352, 355, 359, 363, 367, 370, 374, 378, 382, 386, 390, 426.

<sup>4</sup> Compl. ¶¶ 169, 176-180, 186-190, 196-199, 205-209, 215-219, 225-229, 235-239, 244-247, 253-257, 263-265, 270-274, 279-285, 291-295, 301-305, 311-315, 321-325, 331-335, 341-347, 349, 352, 355, 359, 363, 367, 370, 374, 378, 382, 386, 390.

*OpenAI*.” (*Id.* ¶ 59 (emphasis added).) And Plaintiffs’ contributory infringement allegations state that “Microsoft profited from its [sic] OpenAI’s direct infringement through its investment in OpenAI and its monetization of **GPT-based products**.” (*Id.* ¶ 426 (emphasis added).) From beginning to end, the Complaint solely alleges infringement based on training of OAI’s GPT-based models and, as to Microsoft, for its commercialization of specific products using those models.

Given their claims, Class Plaintiffs must resort to citing paragraphs that refer to allegations regarding “Defendants.” (Dkt. 269 at 3 (citing Compl. ¶¶ 9-10, 393-397, 406, 415).) Nowhere does the Complaint refer to any model other than a GPT-based model or any allegations exclusive to Microsoft’s LLMs, including in the direct infringement allegations aimed at both OpenAI and Microsoft. (Compl. ¶¶ 412-418.) Indeed, the handful of generalized citations on which Class Plaintiffs rely, read properly within the litany of references to GPT, in no way convey that Plaintiffs are also accusing Microsoft of infringement based upon its training of its own LLMs.

Microsoft’s LLMs are simply not at issue in this case. Class Plaintiffs have ways to make them relevant either through an amendment after meeting the appropriate Rule 15 and Rule 16 standards, or through a separate lawsuit. Class Plaintiffs do not have a *de facto* right to add new models to the case through discovery. Discovery in an existing lawsuit cannot properly be used to pursue information in aid of a potential future case. *See, e.g., 01 Communique Lab., Inc. v. Citrix Sys., Inc.*, 2014 WL 25060250, at \*4 n.6 (N.D. Ohio June 3, 2014) (“Nonetheless, a plaintiff is not entitled to accuse certain products of infringement, and then ask for discovery one very other product on a mere suspicion that other products might infringe as well.”).

Class Plaintiffs also cannot rely on their contributory infringement allegations as the basis for wide-ranging discovery into the datasets Microsoft used to train its own LLMs. Microsoft’s training datasets for its own models have nothing to do with Microsoft’s alleged knowledge related to OpenAI’s training of its models. As illustrated in the case cited by Plaintiffs, the knowledge prong of contributory infringement is specific to the actions of the underlying direct infringement claims: “[t]he knowledge standard is an objective one; contributory infringement liability is imposed on persons who know or have reason to know **of the direct infringement**.” *White v. DistroKid*, 2024 WL 3195417, at \*10 (S.D.N.Y. June 24, 2024) (emphasis added). The only alleged direct infringement underlying Plaintiffs’ contributory infringement allegations relates to OpenAI’s training of its LLMs. (Compl. ¶¶ 425-427.) The contents of the datasets Microsoft used to train **Microsoft’s** LLMs are not relevant to whether Microsoft had knowledge of whether **OpenAI** was infringing Plaintiffs’ copyrights via training of its own GPT models.

Any link of this discovery to the willfulness allegations is even more attenuated. As an initial matter, it’s unclear, at best, that Class Plaintiffs have sufficiently alleged willful infringement against Microsoft (an allegation that Microsoft vigorously denies). Class Plaintiffs’ allegations of willful conduct are aimed almost exclusively at OpenAI. (*Id.* ¶¶ 3, 79-130 (“OpenAI’s Willful Infringement of Plaintiffs’ Copyrights”), 175, 185, 195, 204, 214, 224, 234, 243, 252, 262, 269, 278, 290, 300, 310, 320, 330, 340.) This is unsurprising, as the allegations in the Complaint exclusively focus on OpenAI’s training of OpenAI’s LLMs and the Complaint is silent on Microsoft’s training of its own LLMs. In short, there is no rational connection between the data used to train **Microsoft’s LLMs** and any allegation that Microsoft acted willfully in connection with **OpenAI** training its own LLMs.

Accordingly, Plaintiffs’ request for discovery into Microsoft’s (1) training data for its own models and (2) data access strategy for those same models should be denied.

Honorable Ona T. Wang

- 4 -

November 22, 2024

Respectfully submitted,

Respectfully submitted,

*/s/ Annette L. Hurst*

*/s/ Jared B. Briant*

Annette L. Hurst

Jared B. Briant

*Counsel for Defendant Microsoft Corporation*